ISSN 1992-6502 (Print) Вестник УГАТУ ISSN 2225-2789 (Online)
2025. Т. 29, № 4 (110). С. 39-47 http://journal.ugatu.su

УДК 004.85 doi 10.54708/19926502_2025_29411039
Код ГРНТИ 28.23.37

Построение нейросетевого классификатора эмоций для мультимодальных данных

И.С. Косачев*, О.Н. Сметанина, Е.Ю. Сазонова

ФГБОУ ВО «Уфимский университет науки и технологий» (УУНиТ), г. Уфа, Россия

Аннотация: Данное исследование посвящено разработке модели для классификации эмоций человека по мультимодальным признакам. В статье проведен разбор существующих работ, решающих задачу классификации эмоций по голосу и речи; описана постановка задачи классификации эмоций, подготовка данных и методика решения; представлены результаты экспериментов с различными моделями для решения задачи. Для обучения был использован набор данных Dusha, состоящий из аудиозаписей на русском языке. В результате экспериментов была получена модель, объединяющая Wav2Vec2 и DistilBERT-small, которая достигла на тестовом наборе значение f1-macro 0,84 на crowd подвыборке и 0,62 на podcast.

Ключевые слова: *машинное обучение, классификация эмоций, мультимодальные данные, нейронные сети.*

***ilyastalk@bk.ru**

Введение

В настоящее время искусственный интеллект все больше и больше проникает в нашу повседневную жизнь. Одним из таких примеров являются чат боты. Их применяют в различных областях, начиная от консультации при продажах и заканчивая обработкой запросов клиентов колл-центра. Самые простые чат боты являются экспертными системами, в которые заложены ответы на самые популярные вопросы.

Минусом таких ботов является отсутствие гибкости в ответах. При общении с другими людьми мы учитываем не только то, что нам говорит наш собеседник, но и то, как он говорит. Опираясь на его эмоциональное состояние, мы можем эффективнее строить диалог и нам будет проще прийти к взаимопониманию.

Эмоция – психический процесс, отражающий субъективное оценочное отношение человека к различным объектам [1]. Для выражения своих эмоций и передачи с их помощью информации другим людям, человек может использовать вербальные и невербальные средства коммуникации.

Вербальные средства коммуникации представляют собой способ передачи информации посредством речи, с помощью слов. К такому типу относится устная и письменная речь, слушание, чтение. Устная и письменная речь используются для передачи информации, а слушание и чтение для восприятия информации. Невербальная коммуникация, наоборот, включает в себя способы передачи информации без использования слов. К такому типу коммуникации относится мимика, голос, язык телодвижений.

Во время разговора человек передает 40% информации вербально и 60% невербально [1–2]. Получается, что при взаимодействии с клиентом модель теряет 60% информации от клиента. Механизм оценки как вербальных, так и невербальных сигналов может помочь машине лучше взаимодействовать с человеком.

В качестве такого механизма может выступать модель для классификации эмоций по голосу и речи. Благодаря этой модели машина сможет реагировать на интонацию клиента и будет более эффективно взаимодействовать с ним.

Помимо чат-ботов классификация эмоций может потребоваться для рекомендательных систем в различных сферах: в образовании – для оценки эмоционального состояния учеников; при использовании в колл-центрах – для оценки качества предоставляемых клиентам услуг;

в сфере безопасности – для выявления мошенников; в медицине – для определения эмоционального состояния пациентов и др.

Решением задачи классификации эмоций по голосу и речи занимались многие исследователи. Брестер К.Ю., Вишневская С.Р., Семенкина О.Э. [3], Никитин П.В., Осипов А.В., Плешакова Е.С., Корчагин С.А., Горохова Р.И., Гатауллин С.Т. [4] и многие другие посвятили свои работы анализу голоса для классификации эмоций. Плешакова Е.С., Гатауллин С.Т., Осипов А.В., Романова Е.В., Самбуров Н.С. в своих исследованиях [5] искали решение задачи определения эмоциональной тональности текста. Makiuchi M.R., Uto K. и Shinoda K. [6] использовали для классификации эмоции голос и речь человека. Кондратенко В., Соколов А., Карпов Н., Кутузов О., Савушкин Н. и Минькин Ф.Р [7] представили мультимодальный набор данных Dusha, который содержит аудиозаписи на русском языке, размеченные по 4 классам эмоций. Несмотря на множество проводимых исследований для оценки эмоционального состояния человека по наборам данных, в большей части из них модели обучались на англоязычных примерах, из-за чего классификаторы могут хуже работать для русского языка.

Целью данного исследования является разработка модели для классификации эмоции человека по невербальным и вербальным признакам. Для разработки программного обеспечения использовался язык python, и библиотеки PyTorch и transformers – для запуска моделей нейронных сетей. В качестве модели для извлечения текста из аудио использовалась модель whisper, а в качестве модели для классификации использовалась обученная модель, объединяющая Wav2Vec2 и DistilBERT-tiny [1].

Современное состояние проблемы

Проблемой классификации эмоций занимаются многие исследователи. Так, в работе [3] Брестер К.Ю., Вишневская С.Р. и Семенкина О.Э. рассмотрели вопросы распознавания психоэмоционального состояния дистанционного студента по устной речи. Выявленная проблема при работе с аудиозаписями заключается в большом количестве акустических характеристик. Например, они могут иметь низкий уровень вариации, коррелировать друг с другом или содержать зашумленные данные, вследствие чего их использование при распознавании становится нерациональным. Для отбора информативных признаков авторы использовали генетический алгоритм с адаптивной модификацией метода Strength Pareto Evolutionary Algorithm.

Для извлечения признаков из аудиозаписи были использованы программы OpenSmile и Praat. Данные представлены датасетами Berlin, SAVEE и LEGO. В исследовании обозначены модели и методы для классификации: многослойный перцептрон, метод опорных векторов, логистическая регрессия, радиально-базисная нейронная сеть с функцией Гаусса, наивный байесовский классификатор, дерево решений, случайный лес, бэггинг, аддитивная логистическая регрессия, алгоритм генерирования правил 1R. Эффективность моделей сравнивалась на полном и сокращенном после фильтрации наборах признаков. В результате, разработанная авторами [3] методика в большинстве случаев позволила не только повысить качество работы классификатора, но и существенно сократить количество признаков для классификации.

Авторы Плешакова Е.С., Гатауллин С.Т., Осипов А.В., Романова Е.В., Самбуров Н.С.

в работе [5] решали проблему определения тональности текста. В качестве данных для обучения были использованы публикации из социальной сети Twitter на тему «COVID-19». Для векторизации текста рассмотрены классические методы: bag of words и TF-IDF. В основе классификаторов лежат: логистическая регрессия, многослойный персептрон, случайный лес, наивный байесовский метод, метод k-ближайших соседей, дерево решений, стохастический градиентный спуск. В качестве метрик для оценки качества моделей использовали Accuracy, Precision, Recall и F-score. В результате самые высокие метрики были получены с помощью стохастического градиентного спуска с использованием метода bag of words.

Работа [4] Никитина П.В., Осипова А.В., Плешаковой Е.С., Корчагина С.А., Гороховой Р.И. и Гатауллина С.Т. освещает проблему распознавания эмоций по тембру голоса для борьбы с телефонным мошенничеством. Для обучения моделей авторы объединили два набора

данных: TESS и SAVEE. Перед обучением аудиозаписи были преобразованы в мел-кепстральные коэффициенты. На полученных признаках обучены модели: логистическая регрессия, случайный лес, градиентный бустинг, многослойные нейронные сети, сверточные нейронные сети и рекуррентные нейронные сети. В результате самую высокую метрику Accuracy дала сверточная нейросеть.

Makiuchi M.R., Uto K., Shinoda K. [6] проводили классификацию эмоции по речевым и текстовым признакам (по двум модальностям: голосу и тексту). Для каждой модальности использована отдельная модель. Модель для классификации по голосу обучалась выполнению двух задач: реконструированию мел-спектрограммы и классифицированию эмоции. В качестве входных данных модель принимает признаки из Wav2Vec, эмбеддинг идентичности говорящего и последовательность фонем.

Модель для классификации эмоций по тексту состоит из последовательности одномерных сверточных слоев, принимает на вход текстовые признаки, извлеченные из транскрипций фраз с помощью языковой модели BERT. Результатом классификации является взвешенная сумма ответов этих двух моделей. С помощью данного подхода удалось достичь 73% невзвешенной accuracy (UA) и 73,5% взвешенной accuracy (WA) на наборе данных IEMOCAP.

Кондратенко В., Соколов А., Карпов Н., Кутузов О., Савушкин Н. и Минькин Ф. в работе [7] представили открытый мультимодальный датасет Dusha на русском языке. Датасет состоит из двух частей: crowd и podcast. Crowd – часть набора данных, которая была озвучена актерами. В качестве текста авторы решили сгенерировать фразы, которые были бы максимально похожи на реальные запросы пользователей виртуального ассистента «Салют». После этого была произведена псевдо-разметка с использованием модели на основе BERT по 4 классам; злость, радость, грусть и нейтральная эмоция. После этого была проведена оценка псевдоразметки, для этого 10 тысяч фраз разметили вручную и сравнили с разметкой модели. Далее проводилась озвучка текста непрофессиональными актерами. Podcast является набором размеченных вырезок из реальных подкастов. Данная часть набора содержит реальные и спонтанные эмоции, чем значительно повышает разнообразие данных.

Для оценки качества датасета была обучена модель, основанная на архитектуре MobileNetV2, с использованием слоя self-attention. В качестве метрик авторы [7] использовали невзвешенную accuracy (UA), взвешенную accuracy (WA) и макро f1-score (F1). В результате, были получены метрики, описанные в Таблице 1.

Таблица 1 Метрики на тестовом наборе данных.

Набор данных	**Crowd test**			**Podcast test**		
	UA	WA	F1	UA	WA	F1
Dusha	0,83	0,76	0,77	0,89	0,53	0,54

Рассмотренные выше исследования объединяет то, что для решения задачи классификации эмоций они используют машинное обучение, включая как классические подходы [3–5], так и методы глубокого обучения [4, 6, 7]. Авторы работы [6] обучали глубокую нейронную сеть, обученную на мультимодальных данных, что так же показывает применимость машинного обучения для решения данной задачи. Стоит отметить, что в большей части рассмотренных исследований модель обучалась на англоязычных наборах данных [3–6], из-за чего модели могут хуже работать для русского языка.

Постановка задачи

Разработать модель глубокого обучения для классификаций эмоции человека по невербальным и вербальным признакам. Модель должна принимать на входе вербальные и невербальные эмоциональные сигналы человека и возвращать на выходе класс эмоции, которую он выражает. В качестве входных данных выступает аудиозапись, на которой записана человеческая речь. Перед подачей в модель из данной аудиозаписи должны извлекаться вербальные

(текстовая расшифровка речи) и невербальные (интонация, тон голоса и т. д.) признаки, которые будут подаваться в обученную модель машинного обучения.

Математическая постановка задачи может быть представлена в следующем виде. Дано: $X = \{ x_1, x_2, \ldots, x_N\}$ – множество аудиозаписей, содержащих человеческую речь. $Y = \{\overline{y_1}, \overline{y_2}, \ldots, \overline{y_N}\}$ – множество закодированных меток классов, где $\overline{y_i} = \{y'_1, y'_2, \ldots, y'_C\}$ – бинарный вектор размера, C – количество классов (индекс указывает на принадлежность к определенному классу эмоции). $D = \{(x_i, \overline{y_i}) | \, x_i \in X, \overline{y_i} \in Y, i = \overline{1, N}\}$ – обучающая выборка. $F(x)$ – модель для классификации, которая производит отображение $X \to Y$. Требуется для любой пары объектов x_i и $\overline{y_i}$ найти такие параметры модели $F(x)$, чтобы значение целевой функции было минимальной.

В качестве целевой функции будем использовать категориальную кросс энтропию, изображенную на формуле (1). Тогда, задача минимизации будет представлена в следующем виде:

$$CE = -\sum_{i=1}^{N} \overline{y_i} ln(F(x_i)) \to min, \quad (1)$$

где N – количество объектов в обучающей выборке.

Подготовка данных и методика решения

Для обучения модели классификации эмоции будем использовался датасет Dusha [7]. Набор данных содержит аудиозаписи, на которых человек выражает одну из четырех базовых эмоций: злость, радость, грусть и нейтральную эмоцию.

Датасет состоит из двух частей:

1) crowd – набор данных, для которого записи собирались с помощью краудсорсинга;

2) podcast – набор данных, состоящий из фрагментов записей подкастов.

Недостатком датасета является то, что для аудиозаписей поднабора podcast нет текстовых расшифровок. Для их расшифровки была использована модель семейства Whisper-large третьей версии от компании OpenAI. При подаче аудиозаписи в модель дополнительно указывался язык, на котором проговаривается фраза на записи, чтобы повысить качество распознавания.

Помимо аудиозаписи и текстовых расшифровок, для обучения модели методом обучения с учителем важным является наличие меток для каждой записи. Так как разметка проводилась с помощью краудсорсинга, для каждой записи в наборе данных есть несколько меток от разных разметчиков. Для обучения модели требуется для каждой записи произвести агрегацию меток.

В целях упрощения подготовки данных, для обучения модели был использован агрегированный набор меток, который использовали исследователи из Сбера при обучении своей модели [7]. Для агрегации они использовали метод Дэвида-Скина.

Всего производилось три эксперимента:

1) обучение модели градиентного бустинга на признаках из OpenSmile и TF-IDF;

2) обучение мультимодальной нейронной сети на основе двух моделей энкодеров, каждая из которых извлекает информацию из определенной модальности;

3) обучение модели градиентного бустинга на признаках, полученных с помощью обученной мультимодальной модели.

Для проведения первого эксперимента требуется предварительно произвести предобработку текстовых данных и извлечь признаки из аудиозаписей.

Для предобработки текстовых данных использовались библиотеки nltk и pymystem3. С помощью nltk производилась токенизация текста с использованием встроенного WordPunctTokenizer и удаление стоп слов. Далее полученные токены приводились к начальной форме с помощью pymystem3.

Для извлечения признаков из аудио использовалась библиотека OpenSmile. В качестве признаков был выбран набор Geneva Minimalistic Acoustic Parameter Set [8] второй версии, содержащий 88 признаков, среди которых спектральные параметры, частотные характеристики и параметры, описывающие амплитуду.

Проведение экспериментов

Для оценки качества моделей были выбраны метрики accuracy, взвешенная accuracy и f1-macro.

Первым производился эксперимент с обучением градиентного бустинга на признаках из OpenSmile и TF-IDF. В качестве модели градиентного бустинга был использован catboost.

Оценка модели производилась как на всем наборе данных, так и на поднаборах crowd и podcast отдельно. Модель catboost обучалась со стандартными параметрами 600 итераций. Метрики на проверочной выборке и матрица ошибок отображены на Рис. 1.

	all	crowd	podcast
accuracy	0.8024	0.7466	0.9069
accuracy_weighted	0.4979	0.5188	0.3347
f1_macro	0.5671	0.5798	0.3780

Рисунок 1. Метрики модели catboost на признаках из OpenSmile и TF-IDF.

Модель показала очень низкое качество классификации, о чем говорят низкие показатели метрик f1-macro и взвешенной accuracy на всех поднаборах данных. Стоит отметить, что показатели метрики невзвешенной accuracy достаточно высоки и даже достигают значения 0,9 на podcast наборе.

Следующими были проведены эксперименты с дообучением глубокой нейронной сети, состоящей из двух подсетей, которые извлекают признаки из входных модальностей, и полносвязного слоя для классификации.

Аудиозапись проходила следующую предобработку:

1) аудиозапись нормализуется;

2) из середины аудиозаписи извлекается фрагмент длиной 5 секунд (если аудиозапись имеет длину меньше 5 секунд, то наоборот, заполняется нулями для получения нужной длины.

Перед подачей в модель текст разбивается на токены с помощью BPE токенизатора и каждый токен заменяется его индексом в словаре. После полученные последовательности собираются в пакет, в котором они все приводятся к одинаковой длине с помощью добавления нулей.

Архитектура модели изображена на Рис. 2.

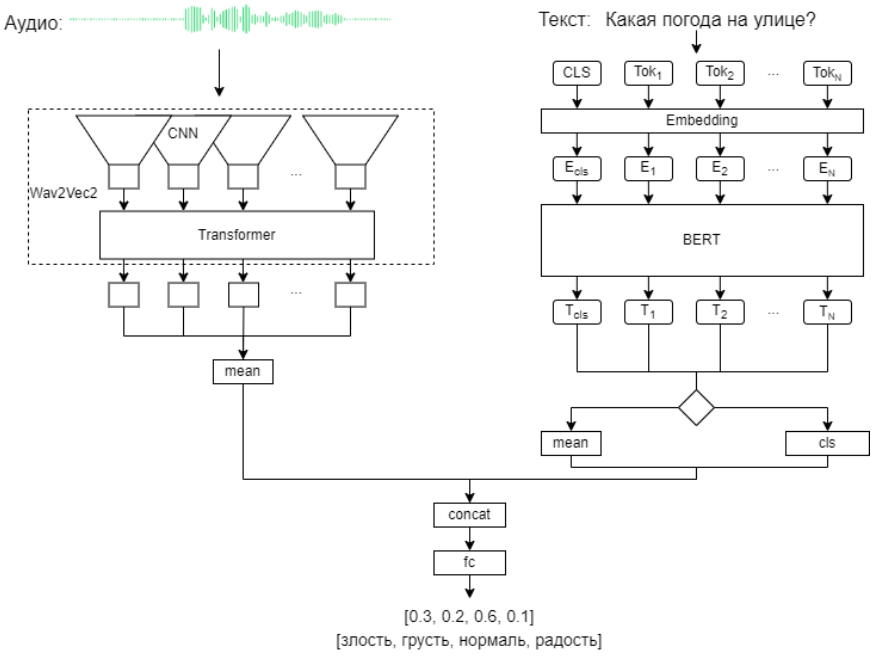

Рисунок 2. Архитектура модели, объединяющей Wav2Vec2 и BERT.

В качестве модели для извлечения признаков из аудио была выбрана Wav2Vec2 [9]. На вход она принимает предобработанную аудиодорожку, а на выход выдает последовательность эмбеддингов. Данная последовательность усредняется, в результате чего на выходе получаем вектор, содержащий информацию о входной аудиозаписи. Такой вектор можно назвать эмбеддингом аудиозаписи. В качестве модели для извлечения признаков из текста использовались модели семейства BERT [10] и DistilBERT [11]. На выход они принимают последовательность индексов токенов, а на выходе выдают последовательность эмбеддингов для каждого токена.

Для того чтобы агрегировать полученные эмбеддинги в один, рассматривается два варианта:

1) извлечь только тот эмбеддинг, который хранит информацию о [CLS] токене (cls агрегация);

2) усреднить все эмбеддинги (агрегация усреднением).

При обучении моделей рассматривались оба варианта агрегации.

После извлечения и агрегирования информации об аудио и тексте, полученные эмбеддинги соединяются и подаются на полносвязный слой для классификации.

Все рассмотренные модели обучались 10 эпох, в качестве оптимизатора был использован AdamW с шагом обучения 0,00001, в качестве функции для изменения шага обучения была использована CosineAnnealingLR, которая за 10 эпох уменьшала значение шага обучения до нуля, в качестве функции ошибки была использована кросс энтропия.

В качестве моделей для извлечения информации из текста рассматривались DistilBERT-tiny и DistilBERT-small от DeepPavlov [12], а также RuBERT от SberDevices [13]. Результаты обучения модели отображены на Рис. 3.

	all			crowd			podcast		
	A	WA	F1	A	WA	F1	A	WA	F1
wav2vec2_distilbert-small_mean	0.8943	0.8160	0.8207	0.8765	0.8358	0.8362	0.9276	0.5798	0.6180
wav2vec2_distilbert-small_cls	0.8926	0.8222	0.8194	0.8748	0.8388	0.8354	0.9260	0.6176	0.6322
wav2vec2_distilbert-tiny_cls	0.8934	0.7596	0.8051	0.8736	0.7866	0.8216	0.9308	0.5047	0.5824
wav2vec2_rubert_cls	0.8910	0.7485	0.8002	0.8691	0.7740	0.8153	0.9319	0.5093	0.5860
wav2vec2_rubert_mean	0.8887	0.7419	0.7943	0.8690	0.7741	0.8142	0.9257	0.4523	0.5278
wav2vec2_distilbert-tiny_mean	0.8898	0.7378	0.7941	0.8682	0.7628	0.8097	0.9300	0.4962	0.5759

Рисунок 1. Метрики, полученные в ходе второго эксперимента (A – accuracy, WA – weighted accuracy, F1 – F1-macro).

Наивысшую метрику F1-macro на всем наборе дала модель, в которой для извлечения признаков использовалась DistilBERT-small с агрегацией усреднением. В последнем эксперименте на признаках из этой модели был обучен классификатор на основе catboost. В качестве признаков извлекались объединенные эмбеддинги аудио и текста. Модель классификатора обучалась со стандартными параметрами 600 итераций. Результаты обучения отражены на Рис. 4.

	all	crowd	podcast
accuracy	0.8955	0.8774	0.9290
accuracy_weighted	0.8049	0.8252	0.5783
f1_macro	0.8209	0.8363	0.6206

Рисунок 2. Метрики модели catboost на признаках из модели wav2vec2_distilbert-small.

Результаты экспериментов

В результате было произведено восемь экспериментов. Сводная таблица с отображением все метрик представлена на Рис. 5.

	all			crowd			podcast			weights
	A	WA	F1	A	WA	F1	A	WA	F1	
catboost_distilbert-small	0.8955	0.8049	0.8209	0.8774	0.8252	0.8363	0.9290	0.5783	0.6206	200
wav2vec2_distilbert-small_mean	0.8943	0.8160	0.8207	0.8765	0.8358	0.8362	0.9276	0.5798	0.6180	200
wav2vec2_distilbert-small_cls	0.8926	0.8222	0.8194	0.8748	0.8388	0.8354	0.9260	0.6176	0.6322	200
wav2vec2_distilbert-tiny_cls	0.8934	0.7596	0.8051	0.8736	0.7866	0.8216	0.9308	0.5047	0.5824	104
wav2vec2_rubert_cls	0.8910	0.7485	0.8002	0.8691	0.7740	0.8153	0.9319	0.5093	0.5860	272
wav2vec2_rubert_mean	0.8887	0.7419	0.7943	0.8690	0.7741	0.8142	0.9257	0.4523	0.5278	272
wav2vec2_distilbert-tiny_mean	0.8898	0.7378	0.7941	0.8682	0.7628	0.8097	0.9300	0.4962	0.5759	104
catboost_opensmile_tfidf	0.8024	0.4979	0.5671	0.7466	0.5188	0.5798	0.9069	0.3347	0.3780	-

Рисунок 3. Сводная таблица с метриками по всем экспериментам.

На рисунке указаны метрики accuracy (A), взвешенное accuracy (WA), f1-macro (F1) и количество параметров модели (weights). Самыми лучшими моделями оказались wav2vec2_distilbert-small_mean и catboost, обученная на признаках из этой модели. Для сравнения с существующими решениями будем использовать wav2vec2_distilbert-small_mean, поскольку, хоть она и незначительно уступает модели catboost по f1-macro, она обходит ее по взвешенному accuracy на всем наборе данных. Сравнение разработанной модели и существующих решений представлено в Табл. 2.

Таблица 2. Сравнение разработанной модели и существующих решений.

	crowd			podcast		
	A	WA	F1-macro	A	WA	F1-macro
Dusha baseline (MobileNetV2 + Self-Attention) [7]	0,83	0,76	0,77	0,89	0,53	0,54
АБК (TIM-Net)	0,84	0,77	0,78	0,9	0,5	0,55
Wav2Vec2_DistilBERT-small_cls	0,88	0,84	0,84	0,93	0,58	0,62
SberDevices (GigaAM-EMO)	0,9	0,87	0,84	0,9	0,76	0,67

В результате обученная модель смогла обойти по всем метрикам baseline набора данных и решение от АБК. Модель уступила решению от SberDevices по метрикам accuracy и взвешенной accuracy на crowd поднаборе, а также по взвешенной accuracy и f1-macro на podcast поднаборе.

Заключение

В рамках данного исследования была разработана модель глубокого обучения для оценки эмоционального состояния человека по вербальным и невербальным признакам с использованием машинного обучения. Результаты анализа современного состояния проблемы показали применимость методов машинного обучения к решению задачи классификации эмоций на мультимодальных данных. Преимущества набора данных Dusha, имеющего большой размер и представленного на русском языке, позволили использовать его для эксперимента.

Проведенные экспериментальные исследования по предложенной методике подготовки данных и оценке эмоционального состояния позволили выбрать наиболее подходящую модель по качеству классификации, а именно модель, объединяющую Wav2Vec2 и DistilBERT-small с агрегацией усреднением. Выбор модели осуществлен по метрикам с лучшими значениями f1-macro: 0,84 на crowd наборе и 0,62 на podcast наборе.

В дальнейшем планируется развивать данное направление исследования путем добавления новых модальностей, а также с помощью обучения модели, устойчивой к помехам, например, в случае неразборчивой речи.

Литература:

1. Косачев И. С. Программное обеспечение для классификации эмоций человека по вербальным и невербальным признакам // Мавлютовские чтения: Материалы XVIII Всероссийской молодежной научной конференции. 2024. Т. 5. С. 265–270. [Kosachev I.S. Software for classifying human emotions by verbal and nonverbal characteristics // Mavlutov Readings: Proceedings of the 18th All-Russian Youth Scientific Conference. 2024. Vol. 5. P. 265–270 (in Russian)].
2. Пиз А., Пиз Б. Новый язык телодвижений. М.: Эксмо. 2012. 400с. [Pease A., Pease B. New body language. Moscow: Eksmo Publishing House. 2012. 400 p. (in Russian)].
3. Брестер К. Ю., Вишневская С.Р., Семенкина О.Э. Распознавание психоэмоционального состояния дистанционного студента по устной речи адаптивными интеллектуальными информационными технологиями // Вестник Сибирского государственного аэрокосмического университета им. академика М.Ф. Решетнева. 2014. № 3(55). С. 35–41. [Brester K. Yu., Vishnievskaya S.R., Semenkina O.E. Speech-based emotion recognition of the distant student with adaptive intellectual information technologies // Vestnik of Siberian State Aerospace University named after Academician M.F. Reshetnev. 2014. No. 3(55). P. 35–41 (in Russian)].
4. Никитин П. В., Осипов А. В., Плешакова Е. С., Корчагин С.А., Горохова Р.И., Гатауллин С.Т. Распознавание эмоций по аудио сигналам как один из способов борьбы с телефонным мошенничеством // Программные системы и вычислительные методы. 2022. № 3. С. 1–13. [Nikitin P.V., Osipov A.V., Pleshakova E.S., Korchagin S.A., Gorokhova R.I., Gataullin S.T. Emotion recognition by audio signals as one of the ways to combat phone fraud // Software Systems and Computational Methods. 2022. No. 3. P. 1–13 (in Russian)].
5. Плешакова Е. С., Гатауллин С. Т., Осипов А. В., Романова Е.В., Самбуров Н.С. Эффективная классификация текстов на естественном языке и определение тональности речи с использованием выбранных методов машинного обучения // Вопросы безопасности. 2022. № 4. С. 1–14. [Pleshakova E.S., Gataullin S.T., Osipov A.V., Romanova E.V., Samburov N.S. Effective classification of natural language texts and determination of speech tonality using selected machine learning methods // Security Issues. 2022. № 4. P. 1–14. (in Russian)].
6. Makiuchi M.R., Uto K., Shinoda K. Multimodal Emotion Recognition with High-Level Speech and Text Features. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2021. P. 350–357.
7. Kondratenko V., Sokolov A., Karpov N., Kutuzov O., Savushkin N., Minkin F. Large Raw Emotional Dataset with Aggregation Mechanism. ArXiv, 2022.
8. Eyben F., Scherer K. R., Schuller B.W., Sundberg J., André E., Busso C., Devillers L.Y., Epps J., Laukka P., Narayanan S.S., Truong K.P. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing // IEEE Transactions on Affective Computing. 2016. Vol. 7. No. 2. P. 190–202.
9. Baevski A., Zhou H., Mohamed A., Auli M. Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NeurIPS, 2020. P. 12449–12460.
10. Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics, 2019.

11. Sanh V., Debut L., Chaumond J., Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv, 2019.
12. Kolesnikova A., Kuratov Y., Konovalov V., Burtsev M. Knowledge Distillation of Russian Language Models with Reduction of Vocabulary. ArXiv, 2022.
13. Zmitrovich D., Abramov A., Kalmykov A., Kadulin V., Tikhonova M., Taktasheva E., Astafurov D., Baushenko M., Snegirev A., Shavrina T., Markov S., Mikhailov V., Fenogenova A. A Family of Pretrained Transformer Language Models for Russian. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. ELRA and ICCL, 2024. P. 507–524.

Об авторах:

КОСАЧЕВ Илья Сергеевич, аспирант второго года обучения Уфимского университета науки и технологий, 32, ул. Заки Валиди, г. Уфа, Республика Башкортостан, 450076, Россия, ilyastalk@bk.ru.
СМЕТАНИНА Ольга Николаевна, доктор технических наук, доцент, профессор кафедры ВМиК Уфимского университета науки и технологий, 32, ул. Заки Валиди, г. Уфа, Республика Башкортостан, 450076, Россия, smoljushka@mail.ru.
САЗОНОВА Екатерина Юрьевна, кандидат технических наук, доцент, доцент кафедры ВМиК Уфимского университета науки и технологий, 32, ул. Заки Валиди, г. Уфа, Республика Башкортостан, 450076, Россия, rassadnikova_ekaterina@mail.ru

Metadata:

Title: Constructing a neural network emotion classifier for multimodal data.
Author 1: Ilya Segreevich Kosachev, second year graduate at the Ufa University of Science and Technology, 32 Zaki Validi st., Ufa, Republic of Bashkortostan, 450076, Russia, ilyastalk@bk.ru, ORCID ID: 0009-0005-2812-6777, Scopus Author ID: 58617515300.
Author 2: Olga Nikolaevna Smetanina, Doctor of Technical Sciences, Associate Professor, Professor of the CMaC Department at the Ufa University of Science and Technology, 32 Zaki Validi st., Ufa, Republic of Bashkortostan, 450076, Russia, smoljushka@mail.ru.
Author 3: Yekaterina Yuryevna Sasonova, Candidate of Technical Sciences, Associate Professor, Associate Professor of the CMaC Department at the Ufa University of Science and Technology, 32 Zaki Validi st., Ufa, Republic of Bashkortostan, 450076, Russia, rassadnikova_ekaterina@mail.ru.
Abstract: This work is devoted to the development of a model for classifying human emotion by multimodal characteristics. The article reviews existing works solving the problem of classification of emotions by voice and speech; describes the setting of the task of classification of emotions, data preparation and solution methodology; presents the results of experiments with different models to solve the problem. Dusha dataset consisting of audio recordings in Russian language was used for the training. The result of the experiments was a model combining Wav2Vec2 and DistilBERT-small, which reached on the test set f1-macro value of 0,84 on the sub-sample crowd and 0,62 on the podcast.
Keywords: machine learning, classification of emotions, multimodal data, neural networks.